ISSN 0120-4157

Biomédica

Revista del Instituto Nacional de Salud

Citación provisional:

Ruano J, Arcila A, Romo-Bucheli D, Vargas C, Rodríguez J, Mendoza O. Deep learning representations to support COVID-19 diagnosis on CT-slices. Biomédica. 2022;42 (2).

Recibido: 18-01-21

Aceptado: 25-11-21

Publicación en línea: 16-12-21

Deep learning representations to support COVID-19 diagnosis on CT-slices

Deep learning to support COVID-19 diagnosis on CT

Representaciones basadas en aprendizaje profundo para dar soporte al diagnóstico del COVID-19 en cortes de TC

Josue Ruano[1], John Arcila[1], David Romo-Bucheli[1], Carlos Vargas[1], Jefferson Rodríguez[1], Oscar Mendoza[1], Miguel Plazas[1], Lola Bautista[1], Jorge Villamizar[1,3], Gabriel Pedraza[1], Alejandra Moreno[1], Diana Valenzuela[2], Lina Vásquez[2], Carolina Valenzuela-Santos[2], Paúl Camacho[2], Daniel Mantilla [2], Fabio Martínez[1]

[1] BIVL[2]ab Biomedical Imaging, Vision and Learning Laboratory, Escuela de Ingeniería de Sistemas e Informática, Universidad Industrial de Santander, Bucaramanga, Colombia

[2] Clínica FOSCAL, Fundación Oftalmológica de Santander, Bucaramanga, Colombia

[3] Facultad de Ingeniería, Universidad de Los Andes, Mérida, Venezuela

Corresponding author:

Fabio Martínez, Escuela de Ingeniería de Sistemas e Informática, Universidad Industrial de Santander, carrera 27 calle 9 oficina 231 edif LP, Bucaramanga, Colombia.

Telephone number: +57 7 6344000 EXT 2110

famarcar@uis.edu.co

Author contributions:

Josue Ruano, David Romo-Bucheli and Fabio Martínez: prepared the first draft and contributed equally to the conception, proposal of the strategy and writing of the final manuscript.

Alejandra Moreno and John Arcila: organized the validation scheme of the databases used, as well as supporting the execution of the proposed experiments.

Jefferson Rodríguez, Oscar Mendoza, Miguel Plazas and Carlos Vargas: development and evaluation of the proposed methodology.

Diana Valenzuela, Lina Vazquez, Carolina Valenzuela, Daniel Mantilla and Paúl Camacho: radiological interpretation, as well as capturing and annotating the radiological studies.

Lola Bautista, Jorge Villamizar and Gabriel Pedraza: implementation and writing of methodology.

Introduction: The coronavirus disease 2019 (COVID-19) has become a significant public health problem worldwide. In this context, CT-scan automatic analysis has emerged as a COVID-19 complementary diagnosis tool allowing for radiological finding characterization, patient categorization and disease follow-up. However, this analysis is dependent on the radiologist expertise, which might result in subjective evaluations.

Objective: To explore deep learning representations, trained from thoracic CT-slices, to automatically distinguish COVID-19 disease from control samples.

Materials and methods: Two datasets were used: SARS-CoV-2 CT Scan (Set-1) and FOSCAL dataset (Set-2). First, the deep representations take advantage of supervised learning models, previously trained on the natural image domain, which are adjusted following a transfer learning scheme. The deep classification was carried out: (a) via end-to-end deep learning approach and (b) via Random Forest and Support Vector Machine classifiers, by feeding the deep representation embedding vectors into these classifiers.

Results: The End-to-end classification achieved an average accuracy of 92.33% (89.70% precision) for Set-1 and 96.99% (96.62% precision) for Set-2. The deep feature embedding with a Support Vector Machine achieved an average accuracy of 91.40% (95.77% precision) and 96.00% (94.74% precision), for Set-1 and Set-2 respectively.

Conclusion: Deep representations have achieved outstanding performance in the identification of COVID-19 cases on CT Scans, demonstrating good characterization of the COVID-19 radiological patterns. These representations

could potentially support the COVID-19 diagnosis on clinical settings.

Keywords: Coronavirus infections/diagnosis; tomography, X-ray computed; deep learning.

Introducción: La enfermedad por coronavirus (COVID-19) es actualmente el principal problema de salud pública en el mundo. En este contexto, el análisis automático de tomografías computarizadas (TC) surge como una herramienta diagnóstica complementaria permitiendo caracterizar hallazgos radiológicos, categorizar y realizar seguimiento de pacientes con COVID-19. Sin embargo, este análisis depende de la experticia de los radiólogos, y las valoraciones pueden ser subjetivas.

Objetivo: Explorar representaciones de aprendizaje profundo entrenadas con cortes de TC torácica para diferenciar automáticamente entre casos COVID-19 y sanos.

Materiales y métodos: Se usaron dos conjuntos de datos de TC: SARS-CoV-2 CT (Conjunto-1) y FOSCAL (Conjunto-2). Primero, modelos de aprendizaje supervisado son previamente entrenados en imágenes naturales, y posteriormente ajustados usando aprendizaje por transferencia. Finalmente, la clasificación se llevó a cabo: (a) usando aprendizaje de extremo-a-extremo, y (b) usando clasificadores como árboles de decisiones y máquinas de soporte vectorial, alimentados por la representación profunda previamente aprendida.

Resultados: El enfoque de extremo-a-extremo alcanzó una exactitud promedio de 92,33% (89,70% de precisión) para el Conjunto-1 y 96,99% (96,62% de precisión) para el Conjunto-2. La máquina de soporte vectorial alcanzó una exactitud promedio de 91,40% (precisión del 95,77%) para el Conjunto-1 y del 96,00% (precisión del 94,74%) para el Conjunto 2.

Conclusión: Las representaciones profundas lograron resultados sobresalientes

caracterizando patrones radiológicos usados para identificar casos de COVID-19 sobre estudios TC, y mostrando ser una potencial herramienta para el soporte del diagnóstico.

Palabras clave: infecciones por coronavirus/diagnóstico; tomografía computarizada por rayos X; aprendizaje profundo.

Coronavirus disease 2019 (COVID-19) emerges nowadays as the major public health problem worldwide due to the third coronavirus outbreak in the last two decades (1,2). According to the Center for Systems Science and Engineering (CSSE), there are 176’349,164 confirmed cases worldwide until June of 2021 (3), while in Colombia there are 3’777,600 confirmed cases and 96.366 deaths associated (3). The COVID-19 is a disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (2,4). This virus belongs to the betacoronavirus genus, which takes a mean of 5 days for incubation, and its manifestation is initially similar to those caused by another respiratory tract virus (2,4). The infection may progress to the lower respiratory with symptoms such as dyspnea with progressive oxygen desaturation, until severe pneumonia, usually present in the second or third week. At advanced stage of this disease, there is a risk of acute respiratory distress syndrome (2,4). This health stage requires specialized clinical intervention and transfer to intensive care units to avoid possible asepsis, septic shock, and even death. Hence, these requirements might overwhelm the public health systems, limiting the adequate provision of services and causing increased mortality in the affected population (5).

The early COVID-19 detection is the most effective strategy to treat and follow patients as well as to decrease disease transmission, allowing quick reactions such as timely lockdowns (6). The gold standard test for COVID-19 diagnosis is the RT-PCR (Reverse Transcription Polymerase Chain Reaction testing) (7), but with a significant report of a high false-negative rate ranging between 20%

to 67% (8). This variance might be caused due to the difficulty of obtaining a high quality sample and the timing of testing (4). Recent work estimating the sensitivity of the RT-PCR on 1,194 inpatients and 1,814 outpatients concludes that it was moderate at best (9). The authors report that when taking into account highly suspicious cases (which never tested positive), the estimated sensitivity (95% CI) was: 67.5% (62.9–71.9%) for inpatients; 34.9% (31.4–38.5%) for outpatients; and 47.3% (44.4–50.3%) for all. Additionally, the delay in the result of the RT-PCR test interferes with an early diagnosis of the disease (10). For those reasons, radiological image analysis has emerged as a powerful technique to support the diagnosis and characterization of symptomatic cases, being a complementary tool in personalized characterization of the disease (11-14). Among others, the analysis of radiological visual patterns over CT scans allows to stratify the disease, to define specific treatments and to follow the evolution from a personalized perspective. In a study presented by Bai *et al.* (15), a group of radiologists with different levels of experience were evaluated in the task of differentiating COVID-19 disease from viral pneumonia on thoracic CT-scans, obtaining a sensibility ranging between 56% and 98%. Nonetheless, the same study showed a low specificity of 25% in radiologists with low experience (15,16). Knowing that the COVID-19 visual patterns on CT-scans are highly similar to other lung infections, experts must go through an arduous training process (15,16). Therefore, the development of computational strategies using radiologic studies to diagnose COVID-19 may help to improve the diagnostic

capacity for health systems and supporting early diagnosis. Additionally, these developments could decrease the high inter-observer variability and the high rate of false-negatives in the COVID-19 detection on CT-scans.

Some artificial intelligence strategies have been developed to accurately diagnose lung diseases using radiologic studies such as pneumonia, pulmonary nodules, chronic obstructive pulmonary disease (COPD), and diffuse pulmonary fibrosis (17,18). Regarding COVID-19 detection on CT-slices, Li *et al*. (5) developed a 3D learning model based on Convolutional Neural Network (CNN) to perform a differential diagnosis between COVID-19 from other lung diseases on thoracic CT-scans (5). Silva *et al*. (19) modified the EfficientNetB0 architecture by adding six layers in the feature extraction stage. In contrast, Ragab and Attallah (20) present a method that combines four CNN and three hand-crafted feature extractors to characterize radiological images exhaustively. Afterward, these features were used to train a Support Vector Machine model (SVM) with the cubic kernel. Both methods applied a transfer learning technique using a deep learning model pre-trained on ImageNet dataset (21). Additionally, the authors used standard data augmentation policies and the evaluation scheme proposed by Soares *et al*. (22). However, the authors do not provide enough information to determine if the evaluation carried out in that paper stratified by patients the training and testing sets. Avoiding such partitioning might result in over-optimistic results, as pointed out by Silva *et al*. (19). Additionally, the low number of cases belonging to different populations and acquisition devices limits the ability of

training generalizable supervised deep learning models.

This work conducted an exploration and analysis of convolutional deep learning representations to support the automatic classification between COVID-19 and non-COVID-19 samples in clinically relevant CT-slices, previously selected by radiologists. From a supervised scheme, a set of architectures originally trained on the natural image domain were adjusted to implicitly identify radiological visual patterns associated to COVID-19. Afterwards, the learned deep representation was used to classify new samples using an end-to-end scheme, but also using the high level embedding vectors with classical machine learning classifiers. This representation was validated on two different datasets separately, showing remarkable results to support the radiological analysis task. The best performance of the proposed strategies yielded scores of 90% (accuracy), 91% (sensitivity), and 94% (specificity) on the mentioned datasets.

Materials and methods

Thoracic CT is an image modality useful to analyze the transverse area, anatomical structure, and density of the lung. Over such images, it is possible to characterize pneumonia disease by a set of radiological findings, as described by the Fleischner Society glossary (23,24). Regarding COVID-19 characterization, there exists some predominant findings that are known to be associated with the disease. Among the most predominant findings, the following can be considered: bilaterally, lower lobe, peripheral and basal predominant ground-glass opacities (GGOs) or consolidation with vascular enlargement (25,26). Also, GGO is superimposed by a mixed pattern composed of

crazy paving, architectural distortion, and perilobular abnormalities (12).

The localization of the radiological findings on thoracic CT-slices are particular for each patient and varies depends on state of disease (11,25,26). Hence, a CT-slice selection process was done for better characterization of the COVID-19 patterns. Such selections were manually performed by radiologists, that explores the whole CT-scan to determine clinically relevant slices. The datasets used in this work are described in the following Subsection.

Datasets

In this work the evaluation of deep learning representation was considered on two different sets, with the main goal to determine the generalization capability of classification, as well as, to determine the effectiveness in retrospective study that counts with demographic patient information. In both cases only axial CT-volumes were considered. The datasets are described as follows:

I. **SARS-CoV-2 CT Scan dataset:** This public collection contains a total of 210 cases comprising 4,173 thoracic CT-slices. A subset of 80 cases are patients infected by SARS-CoV-2 (2168 CT-slices), and 50 cases are non-infected patients (757 CT-slices). The remaining cases are 80 patients with other pulmonary diseases that were not taken into account in this work. For each CT volume, the CT-slice more relevant in terms of radiological findings was manually selected as input of deep learning methodology. This dataset was collected in hospitals of Sao Paulo, Brazil, and all patients were confirmed positive or negative of SARS-CoV-2 by RT-PCR test (22). In consequence, the automated categorization of patients by the deep learning models may be biased to detect

those cases also detected by the RT-PCR test. In our study, this bias is mitigated by the fact that radiological findings identified by the expert were previously selected on the most significant CT-slice for diagnosis.

II. **FOSCAL dataset:** This dataset corresponds to a retrospective study, from which were acquired CT-scans at Clínica FOSCAL, a clinical hospital located in Santander, Colombia, from March 1 to August 19, 2020. The dataset is composed by thoracic CT-scans of a total of 355 patients. A subset of 175 patients were diagnosed positive for COVID-19 infection by RT-PCR (1171 CT slices). The remaining 180 patients were diagnosed negative for COVID-19 (1,364 CT slices), but they could have other pulmonary diseases. Each patient underwent CT and RT-PCR testing for SARS-COV-2. The dataset also contains SARS information in which 1,846 slices do not present it and 416 slices have SARS. Clinically relevant slices were selected by two radiologists with 3 and 4 years of experience. The CT-scans have a spatial variation between [4-15] slices among the tomography. Every single slice has a spatial size of 512 x 512 pixels. The demographic information and comorbidities distribution are shown in table 1.

This research study was conducted retrospectively using human subject data. Approval was granted by the Ethics Committees of Universidad Industrial de Santander and of the FOSCAL clinical center in Bucaramanga, Colombia.

This work introduces deep convolutional representations to deal with COVID-19 automatic classification from thoracic CT slices. These representations aim to recover and learn distinctive visual patterns associated to the disease and properly distinguish among COVID-19 and non-COVID-19 CT images. A transfer-learning scheme was

herein implemented to train and adjust the deep representations in an end-to-end classification setup. As a second alternative for evaluating the deep representation, the last fully connected layers were taken as embedding representations. Then, the embedding representation is used on classical machine learning classifiers such as random forest or support vector machines. The pipeline of this strategy is shown in figure 1.

Convolutional neural network architectures (CNNs)

The CT slice characterization by a CNN is based on a hierarchical representation of the visual patterns distinctive for COVID-19 disease and healthy regions. In general, the first layers of CNN perform a decomposition of the input images, into basic visual primitives. This image decomposition is achieved through a set of kernels learned for the specific task of the convolutional network. Subsequently, more complex patterns are modeled in the upper layers such as relevant texture patterns or regional distributions. For doing so, the activations from the previous decomposition are then convolved with another set of learned filters, which extract patterns of a higher degree of non-linear correlation. Finally, such complex patterns are transformed until reaching a semantic level used as a set of features that represent radiological studies with the presence or absence of COVID-19.

Nowadays, thanks to the success of CNNs on different domains, there exist a wide range of CNNs architectures with particular deep properties and learning specifications (21). In this work, three different CNN architectures were explored. The architectures are conventional yet representative state-of-the-art feature extractors, with promising intermediate representations that capture complex visual representations. The

architectures herein implemented are:

I. **VGG16**: The Visual Geometry Group (VGG) developed a relative very depth network composed of 13 convolutional and 3 fully-connected layers that counts a total of 138 million parameters. This network is characterized to be highly uniform around its layers using multiple stacked small size filters (2 x 2 and 3 x 3) achieving to learn more complex features. This network accomplished first place on ILSVRC 2014 challenge training and testing with the ImageNet dataset of natural images. (27).

II. **ResNet-152**: the Residual Networks consists of a CNN architecture that incorporates identity shortcut connections, which reduces the vanishing gradient problem, creating the so-called residual block. Such connections in the image domain improved the classification performance by training deeper networks than conventional CNN architecture. This network with 60 million parameters won the ImageNet Large Scale Visual Recognition Competition (ILSVRC) in 2015 (28).

III. **InceptionV3**: this net is nowadays one of the most representative architectures, which main proposal lies on reduce the computational cost of deeper networks without affecting generalization through a dimensionality reduction with stacked asymmetric convolutions. First, a 1 x 1 convolution is applied to decrease drastically input dimensions of the large filters. Also, a factorization of such large filters is performed, i.e., an N x N filter is the combination of 1 x N and N x 1 filters. Then, these multiple asymmetric filters were ordered to operate on the same level, get a network progressively wider instead of deeper. With 23 million parameters InceptionV3 shared first place with VGG16 on ILSVRC 2015 (29).

End-to-end classification using Transfer Learning

The CNN architectures used in this work were selected due to its effectiveness on the natural image domain. These representations, however, could be unsuited to represent and differentiate patterns from the radiological domain. Then, an adjusted representation to the radiological problem was herein obtained by using transfer learning. Transfer learning (TL) is a widely known technique that approaches learned weights from large general image representations, and adjusts several layers to an specific domain, CT radiological images in this particular case. Formally, the learned image representation at model Mk (being k the ResNet-152, InceptionV3 or VGG16) is defined as $M_k = \{F, P(F)\}$, being F the feature space and $P(F)$ the marginal probability distribution. In this case, $F = \{f_i(f_{i-1})\}$ represents a hierarchical representation of the general image domain with respect to a particular task, $T = \{D, M\}$. Thus, the task T_t covers the set of classes $D_t = \{d_1, \dots, d_n\}$ defined in the original problem (ImageNet (21)). Then, the idea of TL is to adjust the general codified learning task T_t into a new radiological task T_s, as $T_t = \{D_t, M_t\} \rightarrow T_s = \{D_s, M_s\}$ (30).

Transfer learning ($T_t \rightarrow T_s$) is an adaptive iterative process learned through several epochs, using a relative low learning rate, and using batches of new domain, in this case, trained CT-slices. At the end, a deep representation is obtained for each CNN architecture with the capacity of capturing COVID-19 patterns on thoracic CT-slices.

Classification from pre-trained deep features

A second option to exploit the pre-trained CNN architectures is to use the embedding vectors to represent input CT images. Afterwards, the feature vectors are used to train classical machine learning model such as random forest and support vector machines.

For doing so, the last layer of the CNN nets is flattened into a single vector containing the values associated to different features. The main advantage of this approach is the considerable capacity to characterize complex patterns showing remarkable robustness to distortions, occlusions, and lighting changes (31,32). Additionally, this process reduces the training time and decrease the variability of the results with small datasets (33,34).

Then, two classical machine learning algorithms were applied to the computed vectors for the classification task, a Support Vector Machine (SVM) and Random Forest (RF).

I. **Random Forest**: A RF defines boundaries in the feature space between the COVID-19 and non-COVID- 19 classes. The RF is comprised of a set of independent Decision Tree (DT) algorithms. Each DT is trained over different parts of embedding feature space to reduce prediction variability. A bootstrap aggregating strategy, which consists in randomly selecting a set of training embedding features, was used to build each DT. The final prediction is made by averaging the predictions of the individual trees (35). In this process, we obtain B different trees with the ability to predict the disease y.

II. **Support Vector Machine**: The SVM selects an hyperplane that separates the embedding features of the two classes. The selection is performed by maximizing the distance between the decision limit and the feature vectors from both classes (36). In this work, a polynomial kernel is used to define the decision boundary, because this complex classification problem is not linearly separable. The polynomial kernel is defined as: $k(F_i, F_j) = (1 + \gamma F_i^T F_j)^d$ where F_i and F_j are the $i-th$ and $j-th$ embedded in deep features, d is the degree of the

polynomial kernel and gamma is $1/N$ (37).

Experimental setup

CNNs were previously trained with images from the public ImageNet (21) dataset. The resulting weights were used to initialize a new training process with radiological images. The InceptionV3, Resnet-152, and VGG16 models were trained using a batch size of sixteen (18), and the optimization algorithm was the Adam algorithm. The learning rate was set to 1e 6, while the loss function was binary cross entropy. The presented strategies were evaluated in the two mentioned datasets using a cross-validation setup. Each dataset was split into five folds, and the experiment was carried out independently. For each fold validation experiment, the respective dataset was partitioned with 80% cases for training and 20% for testing. It was also ensured that CT-slices of the same patient were in a single fold, i.e., a patient's CT-slices are contained either in the training or in the testing partition. From such a rule is guaranteed that the model is dedicated to discriminate among pathologies more than associate findings from the same patients. Also, in the experiments carried out in this work considered the same number of slices per patient.

The trained models yield a probability per radiological studies of presence or absence of the disease. A 0.5 probability threshold is used to assign the predicted label 1 (COVID-19) or 0 (non-COVID-19). Then True positives correspond in this work to patients effectively classified with COVID-19. It should be noted that the gold-standard for COVID-19 is based on RT-PCR which has a limited sensitivity for COVID-19 detection. To mitigate potential bias associated with the sensitivity of the RT-PCR test, for each considered CT volume, a CT-Slice selection was carried out by an expert radiologist

based on observed radiological findings. The evaluation of this classification task was measured by first computing the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Afterwards, typical metrics for classification task such as Accuracy (Acc), Sensibility (Sens), Precision (Pre), F1 score. The mentioned metrics are described in table 2. The Area under the receiver operating characteristic curve (AUC) was also computed.

Results

In this section, the performance of the proposed approach was presented separately for each considered dataset and with respect to the two classification schemes.

Performance for SARS-COV-2 CT scans dataset

Table 3 summarize the performance for the three methods using SARS-COV-2 CT scans dataset: a) end-to-end classification from transfer learning approach using VGG16 and ResNet-152 architectures, b) classification from deep features using SVM and RF methods, and c) a baseline strategy presented by Silva *et al*. (19). The baseline strategy also applied a five fold cross-validation scheme using SARS-COV-2 CT scan dataset, separating the 80% of cases (CT-scans by patient) for training and 20% for testing in each fold (19). Other works also used SARS-CoV-2 CT Scan dataset such as Soares *et al*. (22) and Ragab and Attallah (20). However, those are not comparable with the work proposed here because there is no precise information regarding the experimentation setup, as to verify that their training and testing partitions were stratified by cases (patients) as Silva *et al*. suggest (19). It should be also noted the remarkable performance obtained by the VGG16, regarding the accuracy and the AUC metrics. This fact could be associated to the small and dense representation kernels on the first

layers of this net. Also, the results suggest that for the data amount the 16 layers are sufficient to fix a boundary and separate between control and COVID-19 cases.

For the proposed classification method using the deep features (embedding), a fine-tuning was carried out for the RF and SVM classifiers as shown in figure 2. First, RF was tuned by varying the amount of trees in each iteration, taking into account that the maximum depth of the trees is 60. A similar procedure was performed for the Support Vector Machine using a polynomial kernel varying the degrees. The best configuration using RF classifier was for 80 trees obtaining a F1-Score of 93.42%, and for SVM classifier was for 7 degrees of the polynomial kernel achieving a F1-Score of 93.63%. These results show that the embedding classification strategy using an SVM classifier obtained better partitions of the deep feature space, obtaining the highest results to detect CT-slices with COVID-19. All evaluation metrics of the best configurations for both classifiers are shown in table 3.

Previous results (table 3) showed that the embedding method using an SVM classifier achieved the highest performance with an F1-Score of 93.63% and precision of 95.77% outperforming the results obtained by Silva *et al*. (19) which only obtained a high sensibility. Additionally, the embedding method is quite stable across all the folds: the standard deviation of all performance metrics is less than 2.83% in each metric when comparing with the transfer learning method and the baseline.

Performance for FOSCAL dataset

Two methods were evaluated and the performance metrics were computed using the FOSCAL dataset: a) end- to-end classification from transfer learning approach using VGG16, ResNet-152 and InceptionV3 architectures, and classification from deep

features using SVM and RF methods. A similar fine-tuning procedure has been performed for the embedding method over the FOSCAL dataset shown in Figure 3 by using the same parameters in Section 4.1 for SVM and RF classifiers. In this case, the best F1-Score obtained for the SVM classifier with 6 degrees was 96.46%, and for RF classifier with 7 trees was 94.67%.

Table 4 shows the results obtained by the proposed approach using the FOSCAL dataset. An accuracy of 95.57% with a precision of 95.74%, a sensitivity of 95.79% and an F1-Score of 95.57% exhibited that embedding method with SVM classifier provides a better representation of the embedded space and it is able to detect accurately COVID-19 cases on CT-slices in the local population. In this case, the VGG16 was also the best net to represent CT-slices, fact associated to the amount of data used to the transfer learning scheme. At the same time, the other deep nets show remarkable results on the end-to-end representation, achieving in general scores up to 90%.

The evaluation in both dataset shows a remarkable performance of deep representation, which could be key to reduce radiologist subjectivity on analysis and diagnosis over CT-scans. Also, the evaluation over both dataset suggests that best boundaries separation is obtained from embedding vectors with an additional optimization over a SVM hyperparametric space. These embedding vectors recover high semantic level knowledge of deep representation, and the additional non-linear kernel separation could induce to better boundary separation among defined diseases.

Discussion

Nowadays, the main public health problem in the world is the COVID-19 disease, therefore it is fundamental to join forces and establish synergies for innovation and

proposal of alternative and complementary methods that allows the characterization, diagnosis and follow up of this disease. Among others, such efforts may support early diagnosis, mitigate collapse of health services and help with a proper analysis and treatment of more affected patients. This work presented a deep learning representation for COVID-19 detection in thoracic CT-slices. From each CT-scan, an expert selected a set of relevant slides with the most distinctive radiological patterns that represent those infected by COVID-19 or healthy lungs. Hence, deep feature extraction was performed to represent the complex visual patterns of the disease exploring different Convolution Neural Networks. Afterward, an end-to-end learning approach and an embedding classification strategy were evaluated to differentiate COVID-19 cases and non-COVID-19 cases, from such deep features. The three networks used in this work (ResNet-152, VGG16 and InceptionV3) achieved outstanding performance characterizing radiological patterns to detect COVID-19 cases over CT-scans. Such deep features were also used to feed two binary classification frameworks: a) end-to-end learning using different CNN architectures, and b) machine learning approach using SVM and RF models. Finally, these models evaluated new thoracic CT-slices determining whether such image visually corresponds to a lung infected by COVID-19.

The highest performance in the open SARS-CoV-2 CT Scan dataset was achieved by the embedding strategy, even outperforming state of the art methods evaluated in that dataset. On the other hand, the methods were also shown to be capable of identifying positive COVID-19 cases in the FOSCAL dataset. The results obtained by this work shows the potential implementation on clinical routine to support the diagnosis. It should be noted that in both evaluated datasets the positive reference is based on the

RT-PCR test, which may introduce a bias related with false-positive rate of gold-standard. In both datasets, the slice-CT with major information related to radiological findings was selected. This selection process would mitigate the potential bias induced by false negatives. Besides, the computational approach is based on a statistical representation that captures visual patterns from a significant amount of data. Hence, it is expected that trained representation deals with some outliers that result from false-negative annotations.

Currently, several computational strategies have been proposed to detect COVID-19 cases on thoracic CT- slices (19,20,22). Most of these methods have used deep learning based strategies without comparing different network architectures or by using public datasets without any additional information about particular conditions of patients, comorbidities and information related with capture of samples. In contrast, in our work, three of the most representative networks on the state of the art are used to extract deep features. Also the evaluated architectures were evaluated over two different datasets with the main goal of evidencing the capability of models to represent COVID-19 patterns from different acquisition sources. The performance of these deep features in the binary classification task was evaluated using a typical end-to-end approach and classical machine learning models. In order to compare the results obtained in this work and discuss other CNN configurations, the methods presented by Silva *et al*.(19), Ragab and Attallah (20) trained and evaluated using the SARS-CoV-2 CT Scan dataset (22) are also presented and compared to our work.

First, Silva *et al*. (19) proposed a modified EfficientNetB0 architecture. The modified EfficientNetB0 model was initialized using pre-trained weights from the ImageNet

dataset and the newly added layers with normal random values. This model was trained with the original images and the resulting transformations of three data augmentation processes, namely rotation, horizontal flip, and scaling. The quantitative evaluation reported by Silva using SARS-CoV-2 CT Scan dataset were an accuracy of 98.99%, a precision of 99.20%, and sensitivity of 98.80%. The validation scheme proposed by the authors of the SARS-CoV-2 CT Scan dataset (22) does not provide enough information to ensure that the training and validation partitions were stratified by patient. Instead the experimental setup proposed by Silva *et al*. (19) ensures that the training and validation partitions contained different cases (patients). This setup avoids that CT-slices of a particular patient are presented in both partitions. This validation setup yields a more realistic, yet slightly lower, estimation of the performance for Silva *et al*. (19). The authors report an accuracy of 86.6%, precision of 79.7%, and sensitivity of 94.8%. In the proposed approach, an expert selected a set of clinically relevant slices of a CT-scan per patient, and then the partitions set were conformed obtaining an accuracy of 91.40%, precision of 95.77%, sensitivity of 91.58% for the embedding classification approach with an SVM classifier. Our work outperforms the results of Silva *et al*. (19) in two performance metrics. In comparison, the deep feature extraction architectures chosen in Silva's work corresponds to a smaller network (EfficientNetB0 with 5 million parameters) while deeper networks were used in this work (VGG16 and ResNet-152 both have over 60 million parameters). In addition, although the end-to-end learning approach achieved competitive results, the embedding approach found a better boundary to separate the classes obtaining the highest performance.

On the other hand, Ragab and Attallah (20) present a method that combines three

handcrafted and four CNN features. These handcrafted features include the Discrete Wavelet Transform (DWT), gray level co-occurrence matrix (GLCM), and statistical features. And for CNN architectures, this work fused AlexNet, GoogleNet, ShuffleNet, and ResNet-18 features extractors. The resulting feature vector with a size of 6948 fed an embedding approach using an SVM classifier to perform the binary classification. Contrarily to Silva and our work, Ragab and Attallah used the validation scheme proposed by the authors of the SARS-CoV-2 CT Scan dataset (22), which seems to allow that CT-slices of the same patient are both in the validation and training partitions. The results obtained are very high with an accuracy, precision and sensitivity above 99%. Also worth noting, this method used a similar embedding strategy to perform the classification task, but the fused feature vector is highly more complex and computationally more expensive compared with the single CNN model used in this work. Additionally, we evaluated the RF classifier in the embedding workflow with a correct validation scheme achieving the highest precision with 95.62%, outperforming all configurations evaluated in this work.

With the idea of demonstrating that the method is generalizable to different populations and acquisition devices, this work performed an additional evaluation process with data collected locally. The FOSCAL dataset is a collection obtained in Santander, Colombia from different hospitals with diverse CT acquisition devices. The best results were obtained by the end-to-end classification strategy. The strategy yielded a 95.57% accuracy, 95.74% precision, and 95.79% sensibility in the FOSCAL testing set. The end-to-end strategy seems to benefit from the increased number of patients and CT-slices available in the FOSCAL dataset. These results show that the method herein

proposed is able to accurately detect COVID-19 cases using thoracic CT-slices from two different populations demonstrating to be competitive with other studies on the state of the art. The proposed strategy is nonetheless dependent on the selection of a significant CT-slice which may limit the automatic detection framework. Moreover, in this analysis some additional slices with complementary information about radiological findings are discarded. This fact may be a limitation of the proposed approach that could include additional information to better discriminate COVID-19 patterns with respect to other classes.

Future work includes training and evaluating this method in a cross-dataset setup to ensure the proper COVID-19 disease detection in a larger set of images acquired via CT imaging. In addition, an automatic selection procedure for the clinically relevant CT-slices might be a useful tool to facilitate the integration of the proposed strategy in clinical routine practice. In such sense, the use of an additional stratification related to the stage of the disease could be useful to build and re-train models with more discriminative information. Moreover, the exploration of new deep alternatives may be useful to process the complete CT-volumes. In fact, in the literature today there exist 3D convolutional nets that could be considered in future perspectives to try the problem of automatic COVID-19 diagnosis.

Conflict of interest

None.

Funding

This research work was funded by a grant from Ministerio de Ciencia, Tecnología e Innovación of Colombia (MINCIENCIAS) during the conduct of the project “Sistema de

aprendizaje profundo automático para la identificación temprana y seguimiento de pacientes con riesgo de síndrome de distrés respiratorio agudo", with code 1102101577294.

References

1. **Guarner J**. Three emerging coronaviruses in two decades: the story of SARS, MERS, and now COVID-19. Am J Clin Pathol. 2020;153:420-1. https://doi.org/10.1093/ajcp/aqaa029
2. **Liu R, Han H, Liu F, Lv Z, Wu K, Liu Y, *et al***. Positive rate of RT-PCR detection of SARS-CoV-2 infection in 4880 cases from one hospital in Wuhan, China, from Jan to Feb 2020. Clin Chim Acta. 2020;505:172-5. https://doi.org/10.1016/j.cca.2020.03.009
3. Johns Hopkins University & Medicine. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). Fecha de consulta: incluir día, mes y año. Disponible en: https://coronavirus.jhu.edu/map.html
4. **Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC**. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. JAMA. 2020;324:782-93. https://doi.org/10.1001/jama.2020.12839
5. **Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, *et al*.** Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology. 2020. https://doi.org/10.1148/radiol.2020200905
6. **Aleta A, Martín-Corral D, Piontti AP, Ajelli M, Litvinova M, Chinazzi M,**

***et al*.** Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. Nat Hum Behav. 2020;4:964-71. https://doi.org/10.1038/s41562-020-0931-9

7. **Wang W, Xu Y, Gao R, Lu R, Han K, Wu G, *et al*.** Detection of SARS-CoV-2 in different types of clinical specimens. JAMA. 2020;323:1843-4. https://doi.org/10.1001/jama.2020.3786
8. **Lauren M. Kucirka, Stephen A. Lauer, Oliver Laeyendecker, *et al***. Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since exposure. Ann Intern Med. 2020;173:262-7. https://doi.org/10.7326/M20-1495
9. Y. **Kortela E, Kirjavainen V, Ahava MJ, Jokiranta ST, But A, Lindahl A, *et al*.** Real-life clinical sensitivity of SARS-CoV-2 RT-PCR test in symptomatic patients. PLoS ONE. 2021;16:e0251661. https://doi.org/10.1371/journal.pone.0251661
10. **Liang LL, Tseng CH, Ho HJ, Wu CY.** Covid-19 mortality is negatively associated with test number and government effectiveness. Sci Rep. 2020;10:12567. https://doi.org/10.1038/s41598-020-68862-x
11. **Inui S, Fujikawa A, Jitsu M, Kunishima N, Watanabe S, Suzuki Y, *et al*.** Chest CT findings in cases from the cruise ship “Diamond Princess” with coronavirus disease 2019 (COVID-19). Radiol Cardiothorac Imaging. 2020;2:e200110. https://doi.org/10.1148/ryct.2020200110
12. **Dai Wc, Zhang Hw, Yu J, Xu Hj, Chen H, Luo Sp, *et al*.** CT imaging and differential diagnosis of COVID-19 Can Assoc Radiol J. 2020;71:195-200.

https://doi.org/ 10.1177/0846537120913033

13. **Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, *et al*.** Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology. 2020;296:E115-7. https://doi.org/10.1148/radiol.2020200432
14. **Hope MD, Raptis CA, Shah A, Hammer MM, Henry TS, *et al*.** A role for CT in COVID-19? What data really tell us so far. Lancet. 2020;395:1189-90. https://doi.org/10.1016/S0140-6736(20)30728-5
15. **Bai HX, Hsieh B, Xiong Z, Halsey K, Choi JW, Tran TML, *et al*.** Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT. Radiology. 2020;296:E46-54 https://doi.org/10.1148/radiol.2020200823
16. **Xiao AT, Tong YX, Zhang S.** False-negative of RT-PCR and prolonged nucleic acid conversion in COVID-19: rather than recurrence. J Med Virol. 2020;92:1755-6. https://doi.org/10.1002/jmv.25855
17. **Mahomed N, van Ginneken B, Philipsen RH, Melendez J, Moore DP, Moodley H, *et al*.** Computer-aided diagnosis for World Health Organization-defined chest radiograph primary-endpoint pneumonia in children. Pediatr Radiol. 2020;50:482-91. https://doi.org/10.1007/s00247-019-04593-0
18. **Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al.** Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122-31. https://doi.org/10.1016/j.cell.2018.02.010

19. **Silva P, Luz E, Silva G, Moreira G, Silva R, Lucio D, *et al*.** COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis. Inform Med Unlocked. 2020;20:100427. https://doi.org/10.1016/j.imu.2020.100427
20. **Ragab DA, Attallah O.** FUSI-CAD: Coronavirus (COVID-19) diagnosis based on the fusion of CNNs and handcrafted features. PeerJ Comput Sci. 2020;6:e306. https://doi.org/10.7717/peerj-cs.306
21. **Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L.** ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition. 2009. https://doi.org/10.1109/CVPR.2009.5206848
22. **Soares E, Angelov P, Biaso S, Froes MH, Abe DK.** SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. medRxiv. 2020. https://doi.org/10.1101/2020.04.24.20078584
23. **Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J.** Fleischner Society: glossary of terms for thoracic imaging. Radiology. 2008;246:697–722. https://doi.org/10.1148/radiol.2462070712
24. **Pan F, Ye T, Sun P, Gui S, Liang B, Li L, *et al*.** Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia. Radiology. 2020;295:715-21. https://doi.org/10.1148/radiol.2020200370
25. **Bernheim A, Mei X, Huang M, Yang Y, Fayad ZA, Zhang N, *et al*.** Chest

CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection. Radiology. 2020;295:200463. https://doi.org/10.1148/radiol.2020200463

26. **Parekh M, Donuru A, Balasubramanya R, Kapur S.** Review of the chest CT differential diagnosis of ground-glass opacities in the COVID era. Radiology. 2020:297:E289-302. https://doi.org/10.1148/radiol.2020202504
27. **Simonyan K, Zisserman A.** Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. Fecha de consulta: incluir día, mes y año. Disponible en: https://arxiv.org/abs/1409.1556
28. **He K, Zhang X, Ren S, Sun J.** Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition. 2016. https://doi.org/10.1109/CVPR.2016.90
29. **Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z.** Rethinking the inception architecture for computer vision. IEEE Conference on Computer Vision and Pattern Recognition. 2016. https://doi.org/10.1109/CVPR.2016.308
30. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A Survey on Deep Transfer Learning. In: Kůrková V., Manolopoulos Y., Hammer B., Iliadis L., Maglogiannis I, editors. Artificial Neural Networks and Machine Learning – ICANN 2018. Lecture Notes in Computer Science, vol 11141. Springer, Cham: falta ciudad de publicación; 2018. https://doi.org/10.1007/978-3-030-01424-7_27

31. **Hijazi S, Kumar R, Rowen C.** Using convolutional neural networks for image recognition. San Jose, CA, USA: Cadence Design Systems Inc; 2015. p. 1–12.
32. **Akilan T, Wu QJ, Jiang W.** A feature embedding strategy for high-level CNN representations from multiple convnets. IEEE Global Conference on Signal and Information Processing. 2017. https://doi.org/10.1109/GlobalSIP.2017.8309150
33. **Pham TD.** A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks. Sci Rep. 2020;10:1–8. https://doi.org/10.1038/s41598-020-74164-z
34. **Alebiosu DO, Muhammad FP.** Medical Image Classification: A Comparison of Deep Pre-trained Neural Networks. IEEE Student Conference on Research and Development. 2019. https://doi.org/10.1109/SCORED.2019.8896277
35. **Breiman L.** Random forests. Machine Learning. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324
36. **Savas C, Dovis F.** The impact of different kernel functions on the performance of scintillation detection based on support vector machines. Sensors. 2019;19:5219. https://doi.org/10.3390/s19235219
37. **Amami R, Ayed DB, Ellouze N.** Practical selection of SVM supervised parameters with different feature representations for vowel recognition. arXiv:150706020. Fecha de consulta: incluir día, mes y año. Disponible en: https://arxiv.org/abs/1507.06020

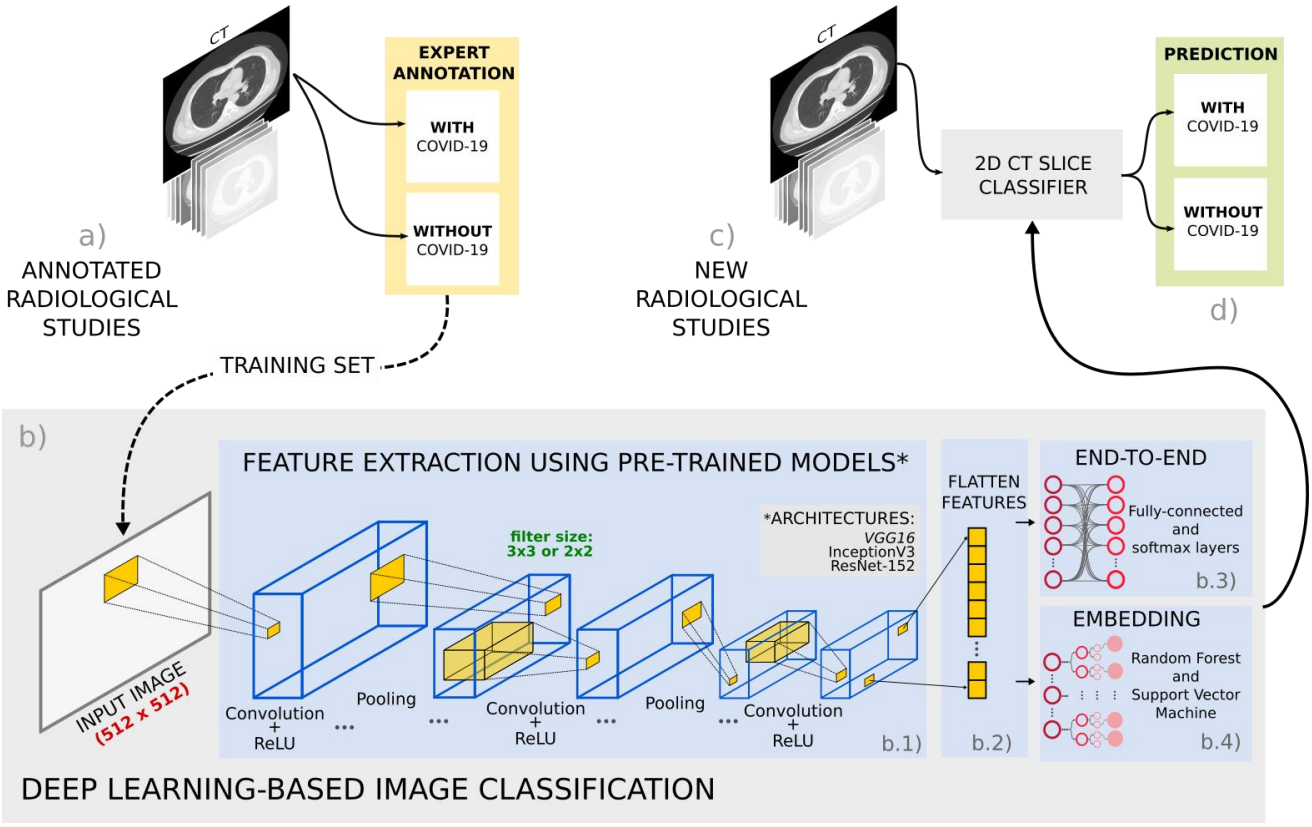

Figure 1. Pipeline of the proposed approach. (a) First, a set of radiological studies were collected from different databases with expert annotations. (b) Afterwards, a deep learning based strategy was trained to detect COVID-19 cases in three steps: (b.1) Different convolutional neural network architectures were tested to characterize the radiological studies. (b.2) Subsequently, the extracted features were flattened to be used as input for the two proposed classification stages: (b.3) end-to-end approach with fully-connected layer classifier, and (b.4) embedding approach with machine learning classifiers. (c) At testing stage, new radiological studies are labeled as with or without COVID-19 using the trained models.

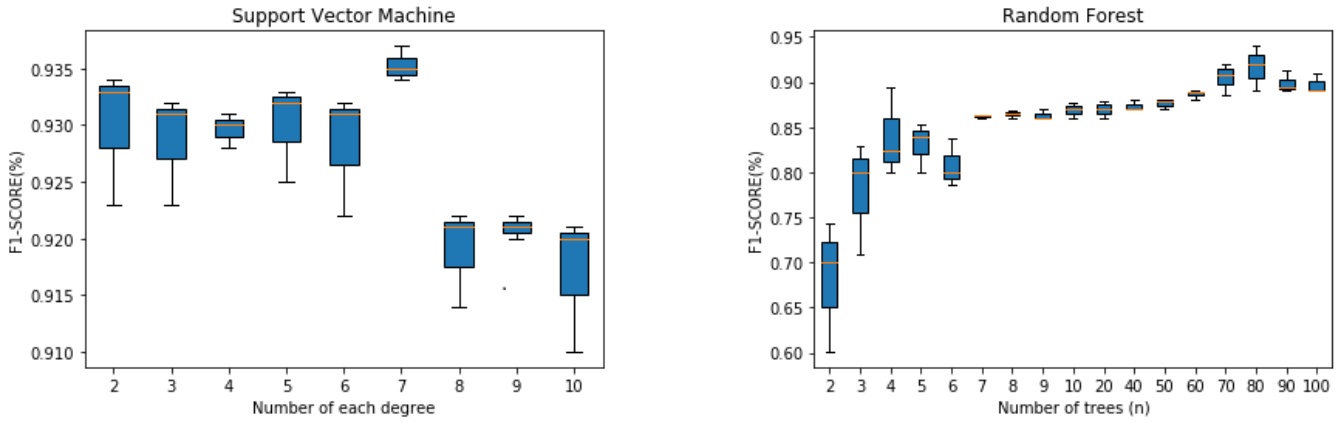

Figure 2. SARS-CoV-2 dataset average results of embedding with Random Forest and Support Vector Machine

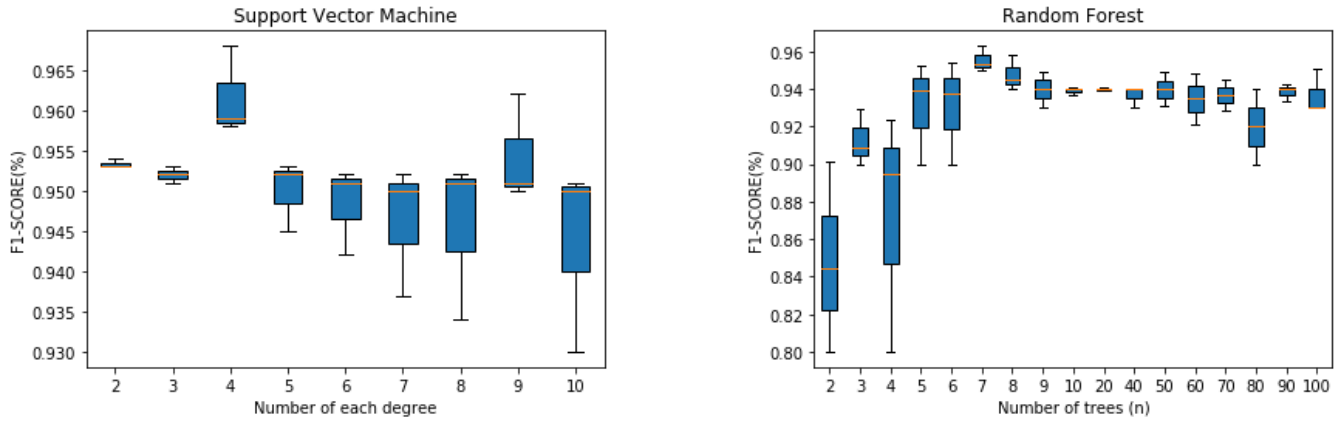

Figure 3. FOSCAL dataset average results of embedding with Random Forest and Support Vector Machine.

Tables

Demographic characteristics	Classes	
	COVID-19	**Non-COVID-19**
Number of patients	175	180
Number of Male/Female/Unknown	109/66/0	68/96/16
Age [range] (mean ± std)	[6 – 92] 60.59 ± 18.68)	[6 – 93] (55.00 ± 17.58)
Comorbidities distribution	46% hypertension 28% no comorbidities 15% cardiovascular disease 11% cancer	59% no comorbidities 28% cancer 7% hypertension 6% others

Table 1. Demographic data and comorbidities distribution of patients included in the FOSCAL dataset.

Accuracy	$Acc = 100 * \frac{TP + TN}{TP + TN + FP + FN}$
Precision	$Pre = 100 * \frac{TP}{TP + FP}$
Sensitivity	$Sens = 100 * \frac{TP}{TP + FN}$
F1 score	$F1 = \frac{2 * TP}{2 * TP + FP + FN}$

Table 2. Metrics used to evaluate the proposed approach. The metrics are based on the quantification of instances: True positives (TP), False positives (FP), True negatives (TN), and False Negatives (FN).

Method	Configuration	Acc (%)	Pre (%)	Sens (%)	F1 (%)	AUC (%)
Silva *et al*. (17)	EfficientNetB0	86.6 ± 10.1	79.7 ± 20.9	94.8 ± 4.50	-	-
End-to-end	VGG16	**92.33 ± 4.81**	89.70 ± 6.74	88.96 ± 6.57	89.89 ± 6.38	**98.20**
	ResNet-152	86.05 ± 1.43	85.52 ± 1.33	76.02 ± 4.01	79.01 ± 3.37	88.51
Embedding	ResNet-152 + RF	90.70 ± 2.80	91.38 ± 2.83	**95.62 ± 2.85**	93.42 ± 2.38	88.82
	ResNet-152 + SVM	91.40 ± 2.48	**95.77 ± 2.83**	91.58 ± 2.41	**93.63 ± 2.80**	91.28

Table 3. SARS-CoV-2 CT Scan dataset average results for the baseline by Silva *et al*. (17), end-to-end and embedding classification approaches.

Method	Configuration	Acc (%)	Pre (%)	Sens (%)	F1 (%)	AUC (%)
End-to-end	VGG16	**96.99 ± 1.10**	**96.62 ± 1.21**	**96.61 ± 1.03**	**96.58 ± 1.11**	**99.50**
	ResNet-152	95.57 ± 5.83	95.74 ± 5.53	95.79 ± 5.52	95.57 ± 5.82	98.87
	InceptionV3	94.11 ± 4.45	94.10 ± 4.46	94.08 ± 4.46	94.07 ± 4.50	98.07
Embedding	ResNet-152 + RF	95.11 ± 2.06	94.81 ± 3.56	95.42 ± 2.96	94.67 ± 2.05	96.06
	ResNet-152 + SVM	96.00 ± 2.56	94.74 ± 2.51	96.00 ± 2.12	96.46 ± 1.84	94.15

Table 4. FOSCAL dataset average results for the end-to-end and embedding classification approaches.